腾讯云

MongoRocks优化与实践

腾讯云-孔德雨

2017.04

目录

MongoRocks简介

- **使用RocksDB作为存储引擎的MongoDB服务**
- **基于由facebook开源的RocksDB**
- **适用于TB级KV-persist业务**
- **在SSD盘上表现优秀,专门为SSD盘的优化**

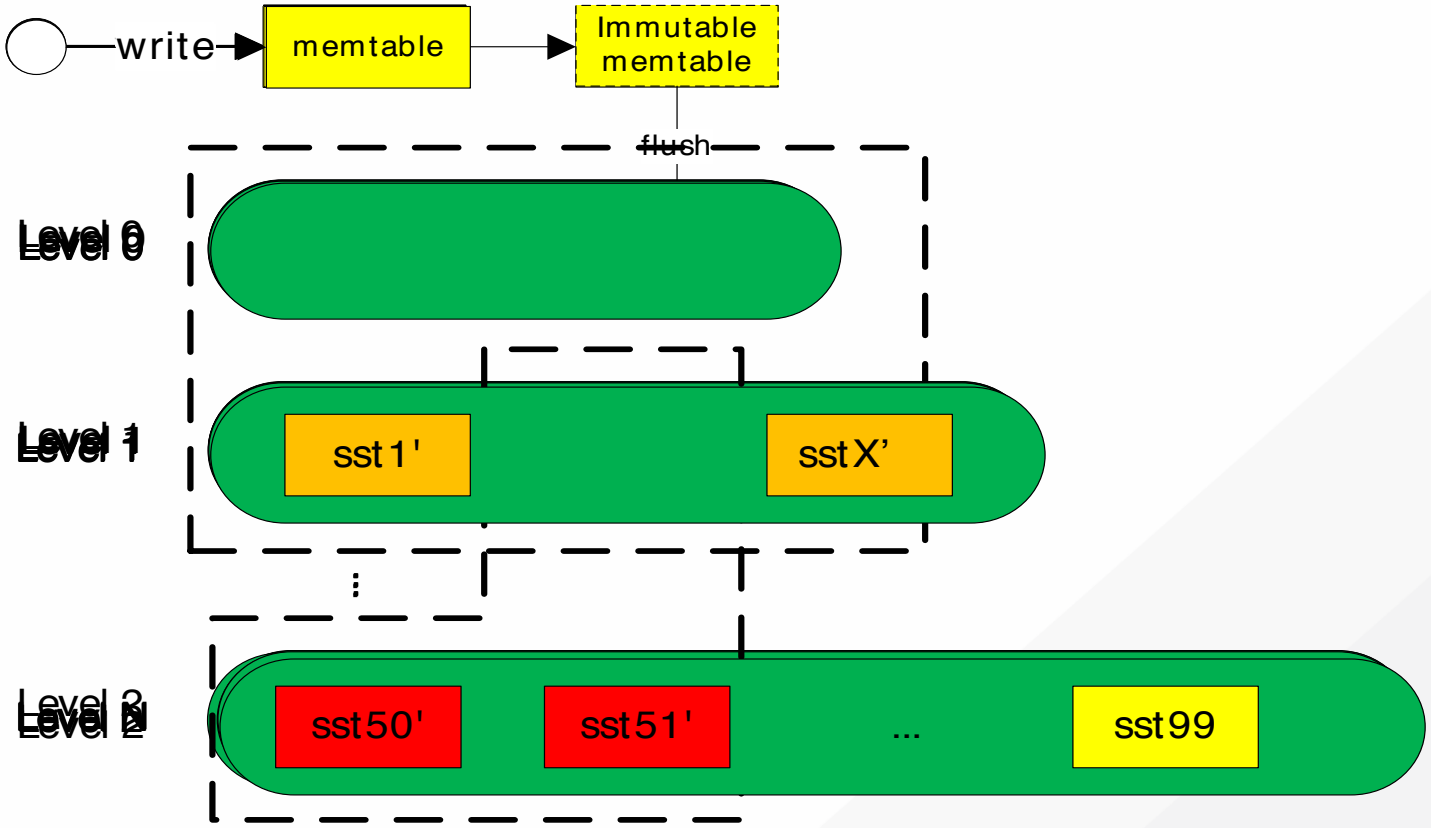
MongoRocks/RocksDB使用情况

- Parse@Facebook(MongoRocks)
- Apache Samza(RocksDB)
- LinkedIn
- 腾讯内部业务(内存数据库的冷数据层)
- 腾讯TRedis(Redis on Rocks)
- TIDB/TIKV

腾讯云MongoDB引擎种类

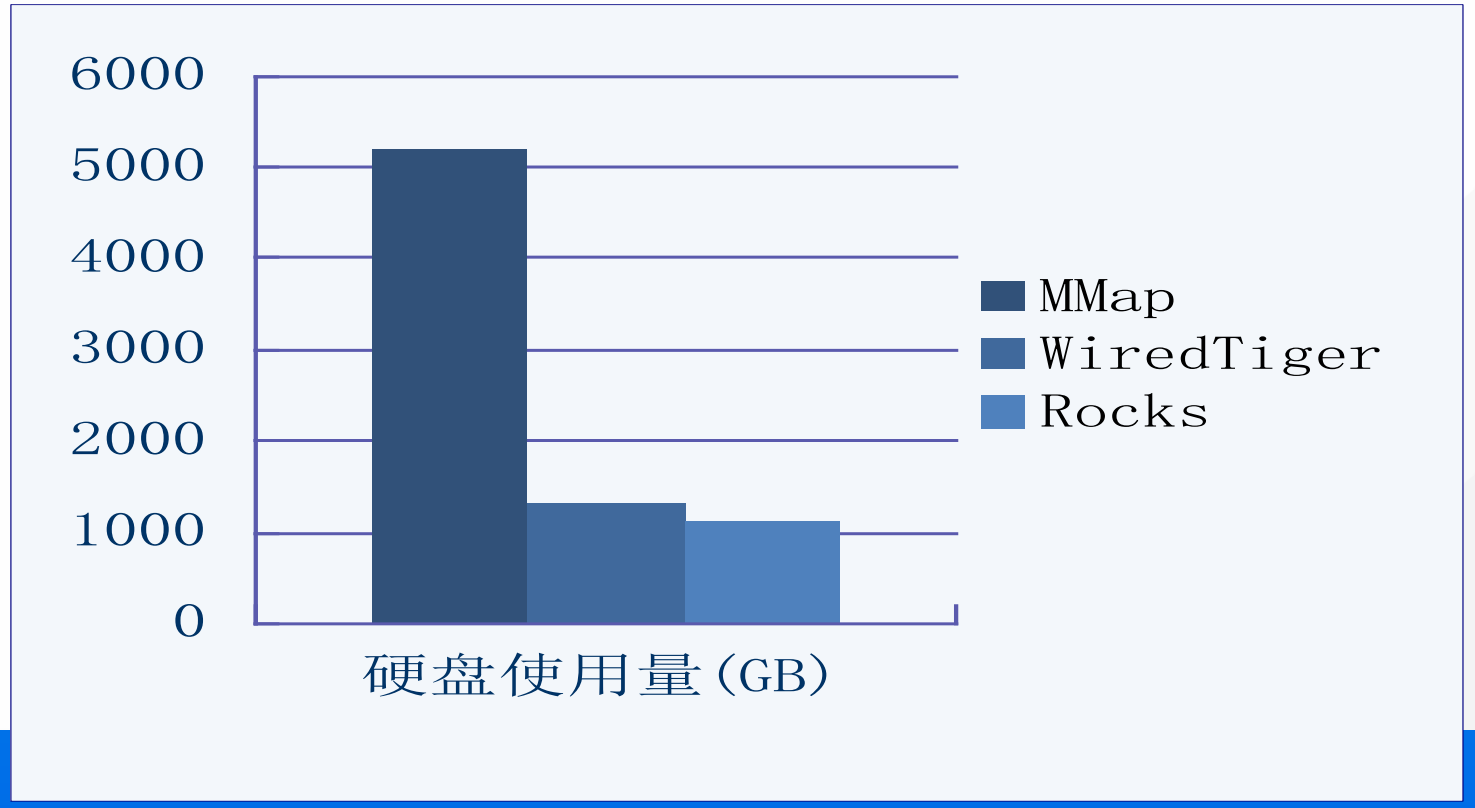
- **得益于MongoDB3.0的插件化引擎接口，目前*腾讯云支持***
  - **WiredTiger(默认支持)**
  - **Mmap(逐步淘汰)**
  - **Rocks(内测用户,国内首家支持MongoRocks的云)**
- **后续计划**
  - **自研引擎/第三方数据库公司合作引擎**

RocksDB原理(LSM)



Rocks对SSD的亲和性

- Rocks将随机写转化为顺序写，提高SSD使用寿命
- Rocks空间利用率高，SSD存储宝贵

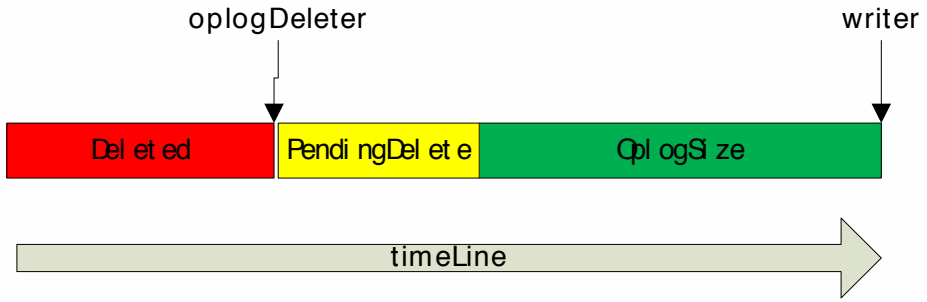


MongoRocks事务性

- MongoDB单行事务性：
  - 索引/数据/Oplog的一致性更新
  - 利用RocksDB::WriteBatch实现
- MongoDB持久性:
  - 基于RocksDB::SyncWAL
  - 支持定期刷盘/每次写入刷盘两种模式
- MongoDB隔离性:
  - SnapShot Isolation 级读写
  - 基于Rocks的快速快照机制, 支持MajorityRead

MongoRocks Oplog大小的维护

- local.oplog.rs的稳定大小:
  - 后台线程标记删除最老的Oplog



MongoRocks对库/表的划分

- MongoRocks默认只有一个ColumnFamily
  - 所有库表放在一个CF中
  - 每一个ns(db.collection)有一个前缀编号
  - 以前缀编号区分数据属于哪个库的哪张表

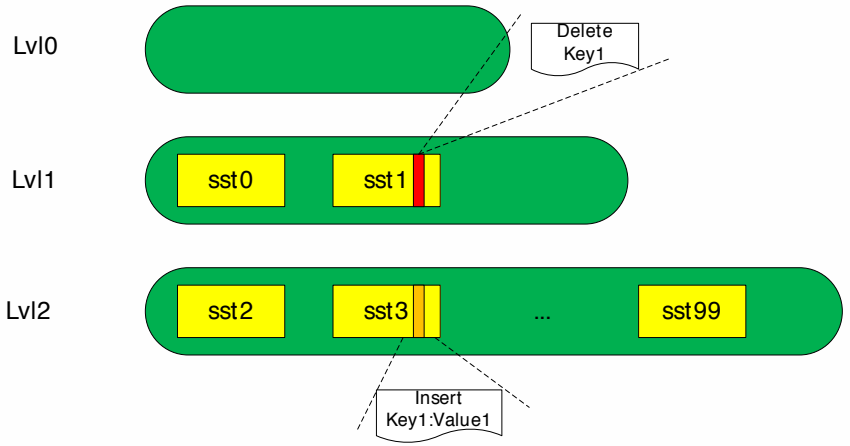
我们对MongoRocks的改进点

分ColumnFamily存储

- 多ColumnFamily
  - kv业务索引少
  - 每个索引单独CF/数据单独CF
- 索引/表快速删除
  - dropColumnFamily 物理删除CF数据
- 便于Oplog按sst文件删除
- 方便按CF为粒度对Cache大小调参(后续PPT)

Oplogs删除策略

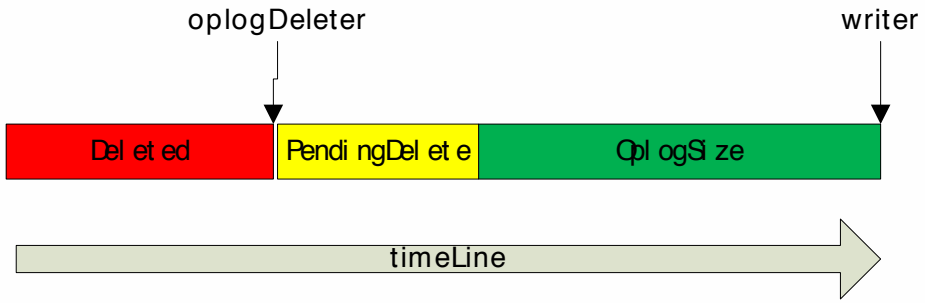
- MongoRocks
  - RocksDB::Delete标记删除(tombStone)



  - 空间不能及时回收
  - tombStone影响查询效率

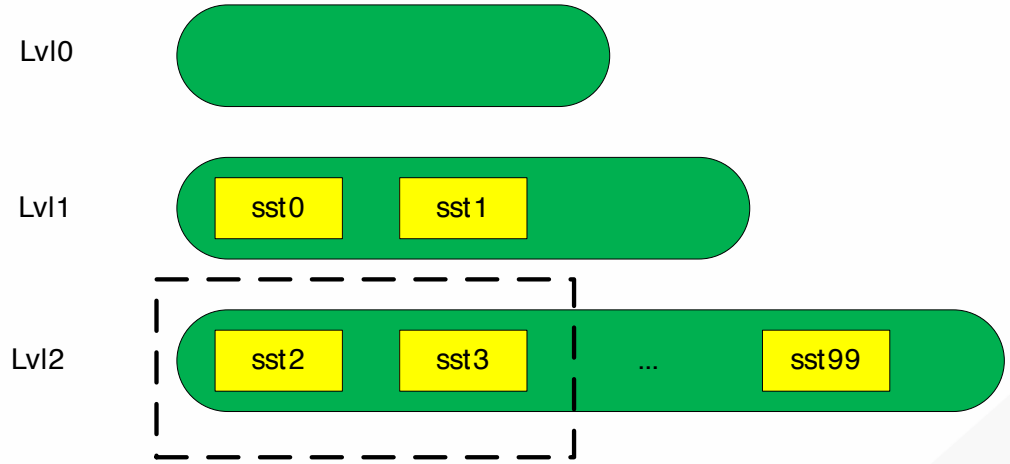
Oplogs删除策略

- Oplog回收的特殊性：
  - Oplog顺序写,无数据交叠
  - 从后往前删除连续的一段Oplog



Oplogs删除策略

- 优化方案:
  - 计算出需要删除的oplog所在的sstfiles
  - RocksDB::DeleteFilesInRange直接按文件删除Oplog
  - 由于分CF存储，不会误删其他表数据

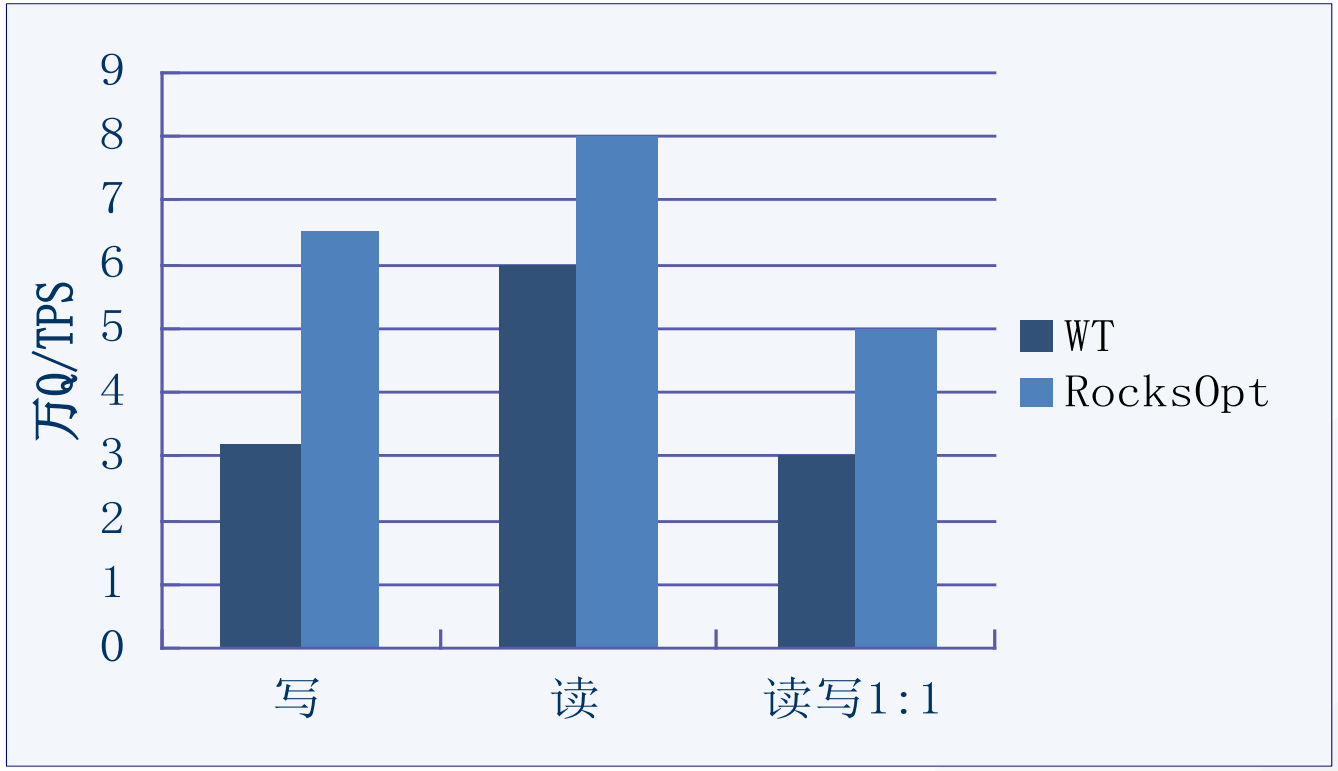


过期数据一般在最后一层的连续sstfiles中

针对KV业务的缓存优化

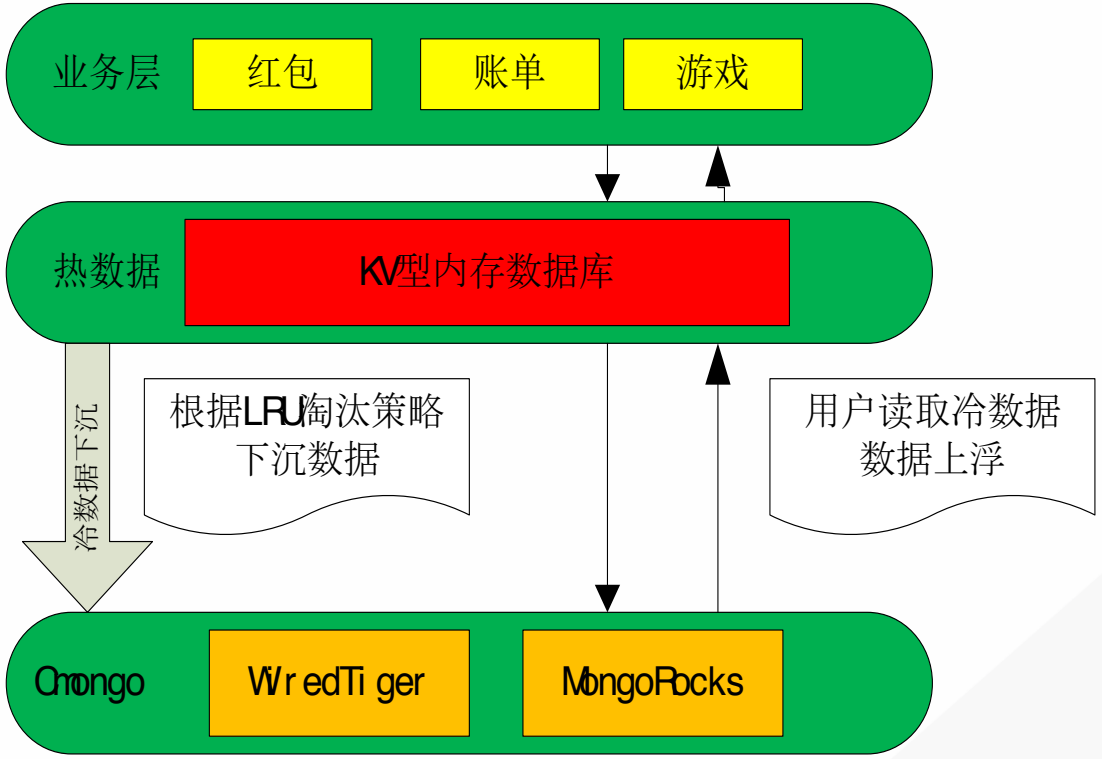
- 开启RowCache，减小BlockCache
  - 对于KV业务,点查询优先于区间查询
- 对于存数据的CF，开启optimize_filters_for_hits
  - 索引CF中存在，数据CF中一定存在
  - 数据CF无bloomFilter的必要性
  - 该参数减小CF的bloomFilter大小

性能对比



腾讯内部MongoDB使用场景举例

- 业务层/冷热数据分离



腾讯云上MongoDB介绍

产品功能

备份&回档
每天业务低峰自动备份，保证数据可靠性，同时为不可预测的业务回档做好准备

平滑扩容
便捷快速的扩容升级功能助力您的业务飞速发展

安全接入和自动容灾
支持私有网络VPC，接入更安全。主节点发生故障，秒级切换备节点，自动容灾

专业监控
多维度监控，随时了解MongoDB的健康情况

产品优势

性能卓越
超大内存 + 高性能SSD组合的物理机型支撑，支持海量访问

集群服务
提供副本集服务，同时您还可以增删节点以满足业务需要

安全可靠
保证数据99.9996%可靠性；至少两份数据备份保障您的数据安全

省心便捷
100%兼容MongoDB协议；多维度监控告警

与自建MongoDB对比

维度	腾讯云数据库MongoDB	自建MongoDB
价格优势	无软硬件投入，提供多种选择（高IO版，容量版）按需付费	硬件：单台存储服务器成本高（如果搭高可用主从（副本集），需要购买2台，资源冗余）软件：需要招聘专业DBA，人力成本高
服务可用性	99.95%，行业高标准，专业团队7*24小时守候，一对一指导，QQ远程协助	需自行处理故障，自建主从，自建RAID
数据可靠性	99.9996%，拥有完善的数据自动备份和无损恢复机制，实时热备，5天内任意时刻数据恢复（注：如两次备份之间操作的数据超过oplog大小，则不可回档至两次备份之间的时间点）	自行保障，依赖硬件的故障发生率，依赖技术人员的数据库管理水平
系统安全性	防DDoS攻击；及时修复各种数据库以及宿主机安全漏洞	自行部署，价格高昂；自行修复数据库安全漏洞
实时监控	多维度监控，故障预警，让您用得安心	需自行开发监控系统，运维人员需半夜处理故障
业务扩容	一键式按需扩容，快速部署，早日上线，让您用得舒心	需自行完成硬件采购，机房托管，应用重新部署等工作，周期较长
资源利用率	按需申请，资源利用率100%，不浪费您一分钱	峰值效用，机器的平均负载不高，资源利用率低

招聘啦！

腾讯云MongoDB核心研发团队招聘(坐标深圳)

我们是谁？
腾讯MongoDB核心研发团队，团队支撑云MongoDB PB-level的存储。核心团队无论开发年限长短，均位于一线coding。

我们做什么？
1）在原生的mongodb基础上重新进行分布式架构的研发。
2）基于云上海量数据，不同业务的瓶颈，优化mongodb的内核，目前主要成果：
a: geo查询的性能优化，在某知名LBS应用厂商的实用场景下，性能有10倍提升。
b: 自研的分布式Mongo架构，以提升数据均衡并发度为出发点，相比3.2版本的数据均衡，性能有5倍以上提升。
c: 针对短连接模型的认证优化，短链接模型下的CPU消耗降低90%。
3）基于nvme盘+MongoRocks和RocksDB的内核开发，以优化kv-persist业务的吞吐量。

我们需要什么样的人？
1. 发自内心的喜欢做技术，有强烈的自我驱动力，具备理性批判的思维，不怕打脸，无偶像崇拜，永不服输。
2. 扎实的技术基本功，关键词：C/C++/linux/数据结构/算法/并发编程。
3. 加分项：数据库理论知识/分布式理论知识/ACM获奖经验/MongoDB(非必要)/RocksDB(非必要）/LevelDB(非必要)。

如何联系我们：
孔德雨 MongoDB研发团队技术负责人
邮箱：deyukong@tencent.com
微信：wolf_kdy